

揭開黑箱模型：探索可解釋人工智慧



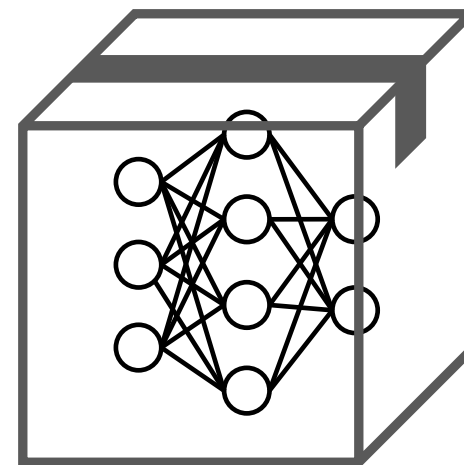
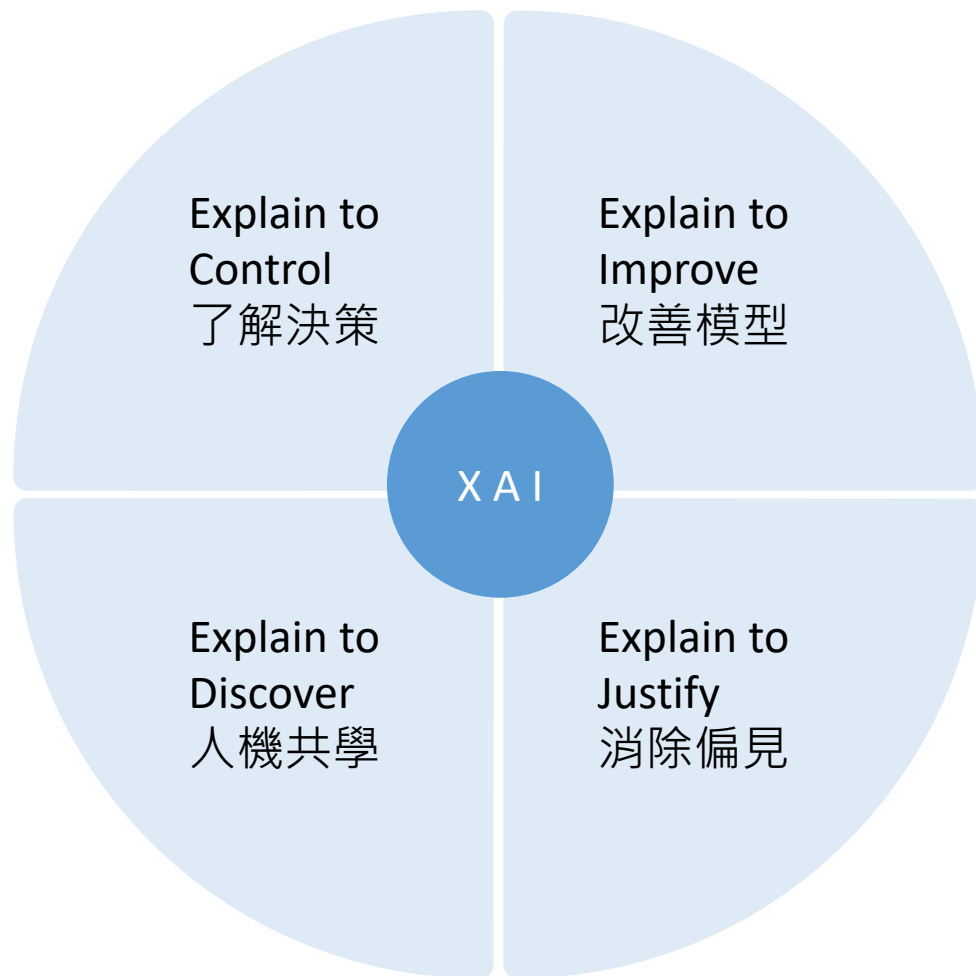
蔡易霖 (10程式中)
iThome 鐵人講堂



10程式中



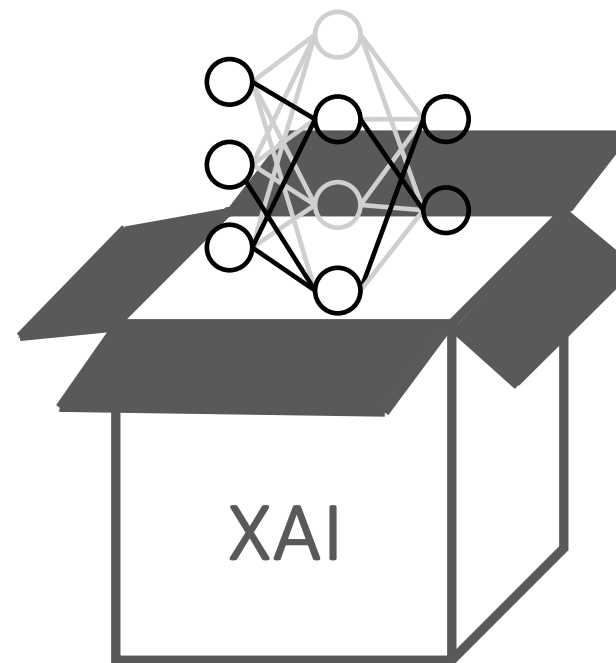
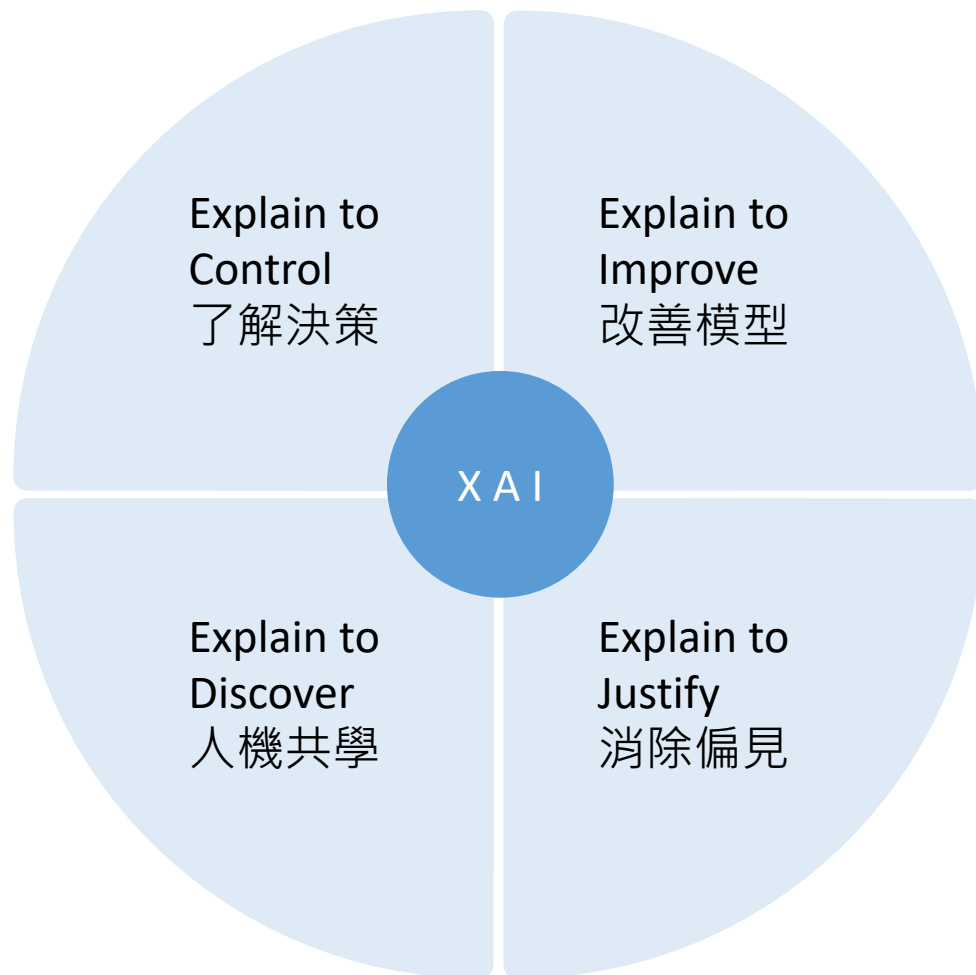
可解釋人工智慧 (Explainable AI)



AI黑盒子



可解釋人工智慧 (Explainable AI)



揭開黑箱模型



AI 模型為何犯錯？

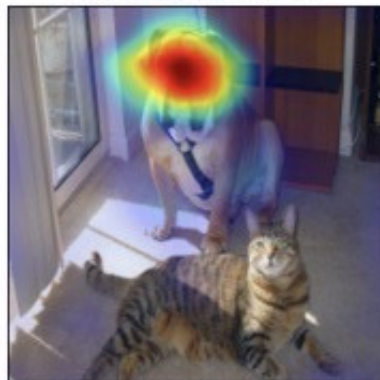
當我們發現模型出錯時，我們通常會想了解其原因。可能的原因如下：

- 資料尚未看過
- 資料看過但學錯

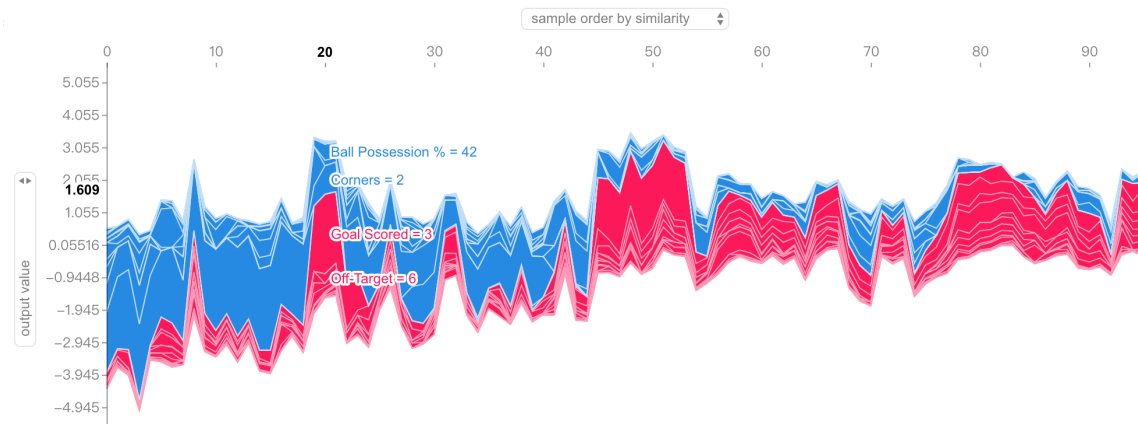
Grad-CAM for "Cat"



Grad-CAM for "Dog"



Grad-CAM



SHAP

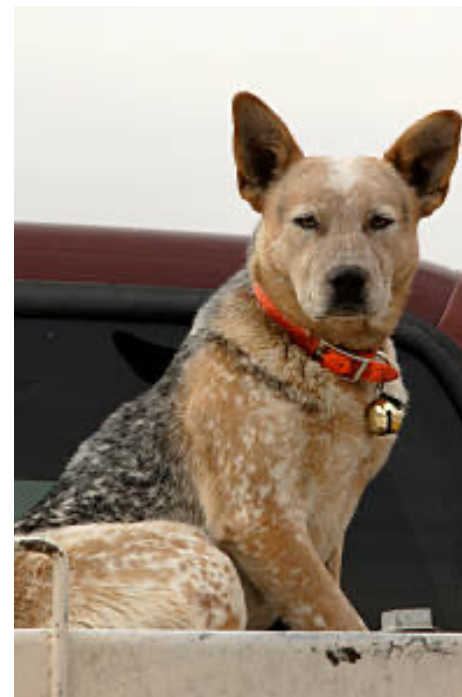


資料看過但學錯



真實答案：貓

AI預測：貓



真實答案：狗

AI預測：貓

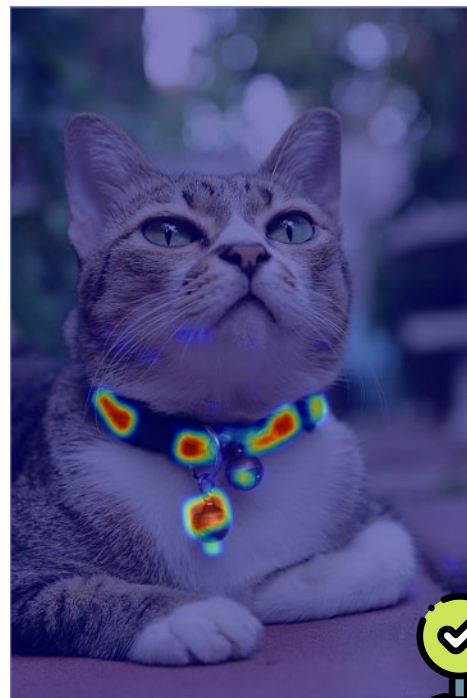


資料看過但學錯

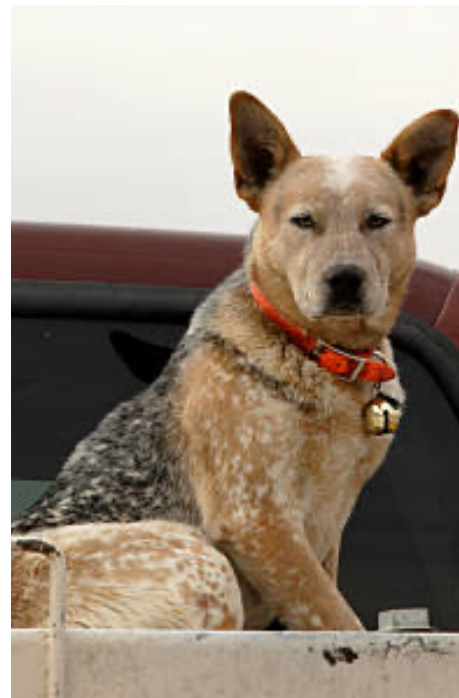
AI判斷貓的規則 => 戴項圈



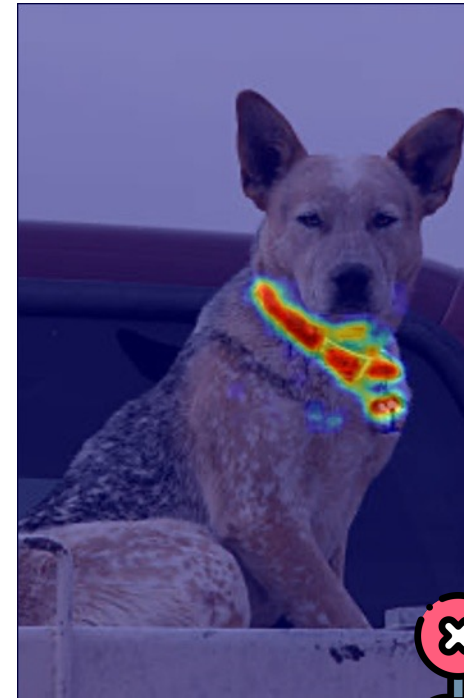
真實答案：貓



AI預測：貓



真實答案：狗



AI預測：貓



並非只有貓咪會戴項圈，狗狗也會戴！



要欺騙AI判讀影像很簡單，只要寫張紙條給他告訴你是誰就行

透過網路上的無限制內容訓練 AI 的結果，也使得 CLIP 出現了人類也會具備的偏見。



Granny Smith	85.6%
iPod	0.4%
library	0.0%
pizza	0.0%
toaster	0.0%
dough	0.1%



Granny Smith	0.1%
iPod	99.7%
library	0.0%
pizza	0.0%
toaster	0.0%
dough	0.0%



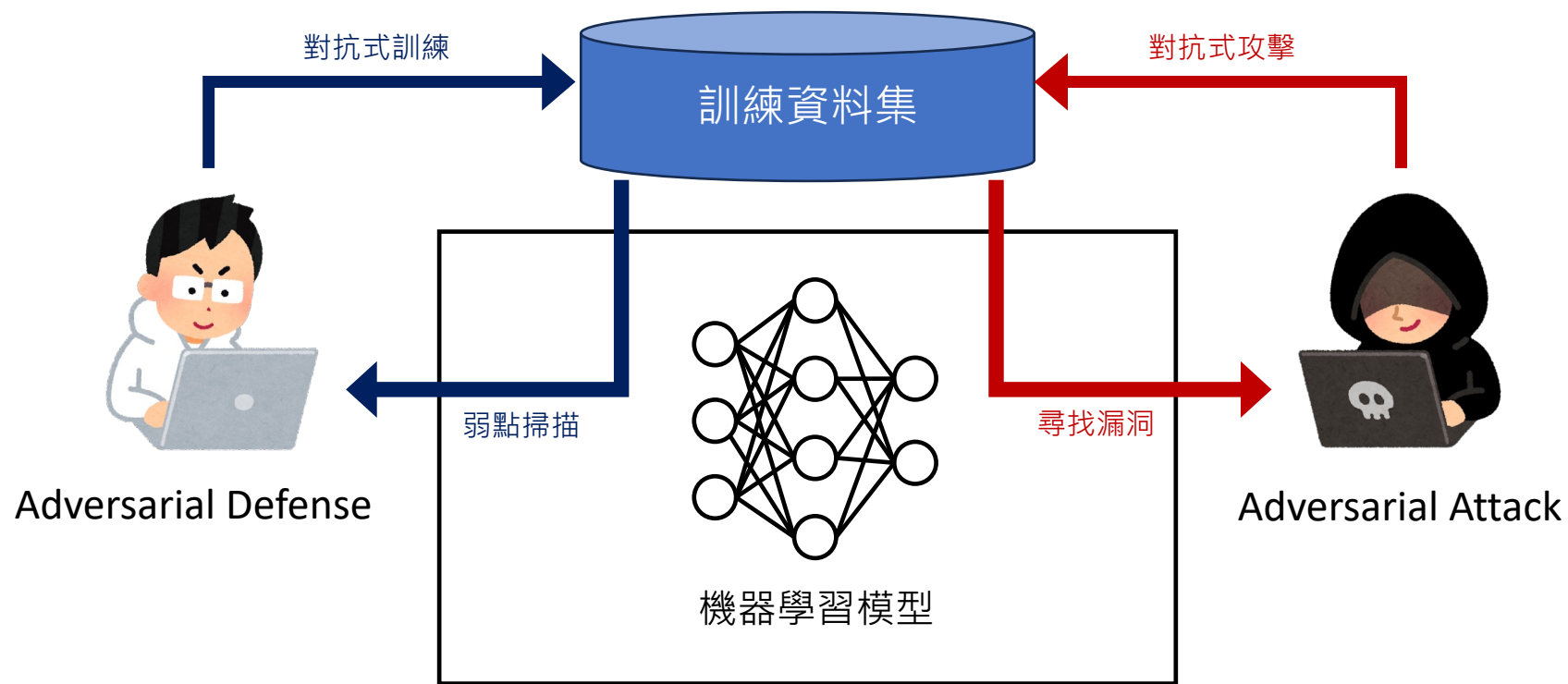
chainsaw	91.1%
lawn mower	7.0%
power drill	1.0%
vacuum cleaner	0.4%
wheelbarrow	0.1%
tractor	0.1%



piggy bank	70.1%
chainsaw	1.5%
slot machine	1.1%
wheelbarrow	0.9%
hammer	0.8%
mousetrap	0.6%

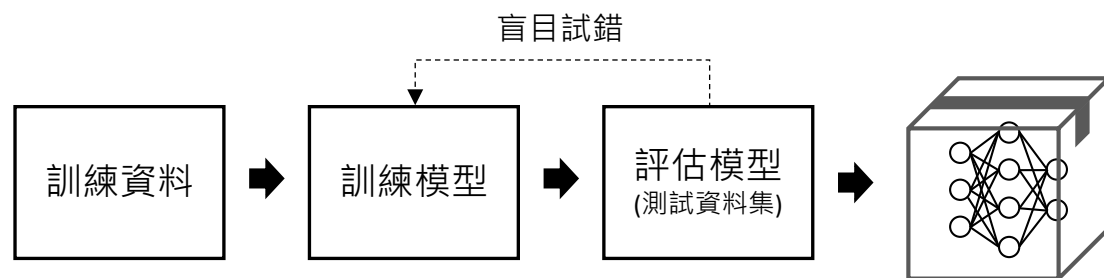
這種明擺著「指鹿為馬」的行為，被研究人員定名為「印刷攻擊」

對抗樣本的挑戰：如何利用XAI檢測模型的弱點？



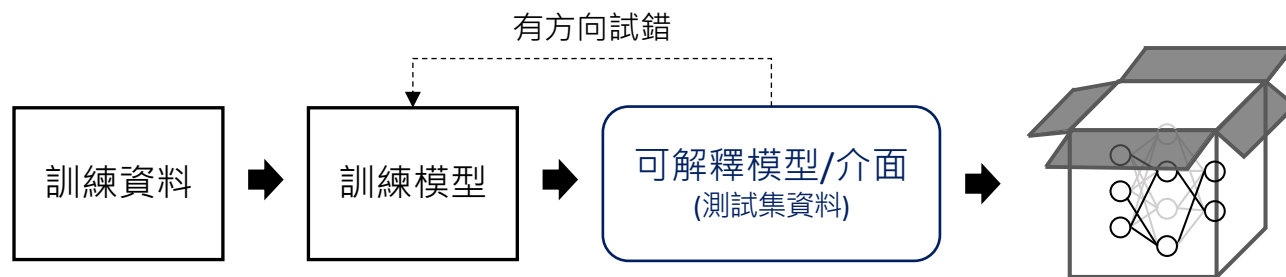
淺談XAI與傳統機器學習的區別

傳統的人工智慧



為何模型這樣做？
為何不選擇其他方法？
什麼時候才會成功/失敗？
何時才能信任模型？
我該如何修正錯誤？

可解釋的人工智慧



我能理解模型為何這樣做？
我明白為何不能這樣做？
我清楚何時會成功/失敗？
我能信任模型推論結果？
我能理解為何模型會犯錯？



各行各業都需要具備可解釋性的AI

商業決策



我可以相信AI的決策嗎？

客服人員



我該如何回覆客人的問題？

製程工程師



我該如何設定最佳的參數？

維運人員



我該如何維運和監控系統？

法庭判決



AI判決是否具有公平性？

科學研發



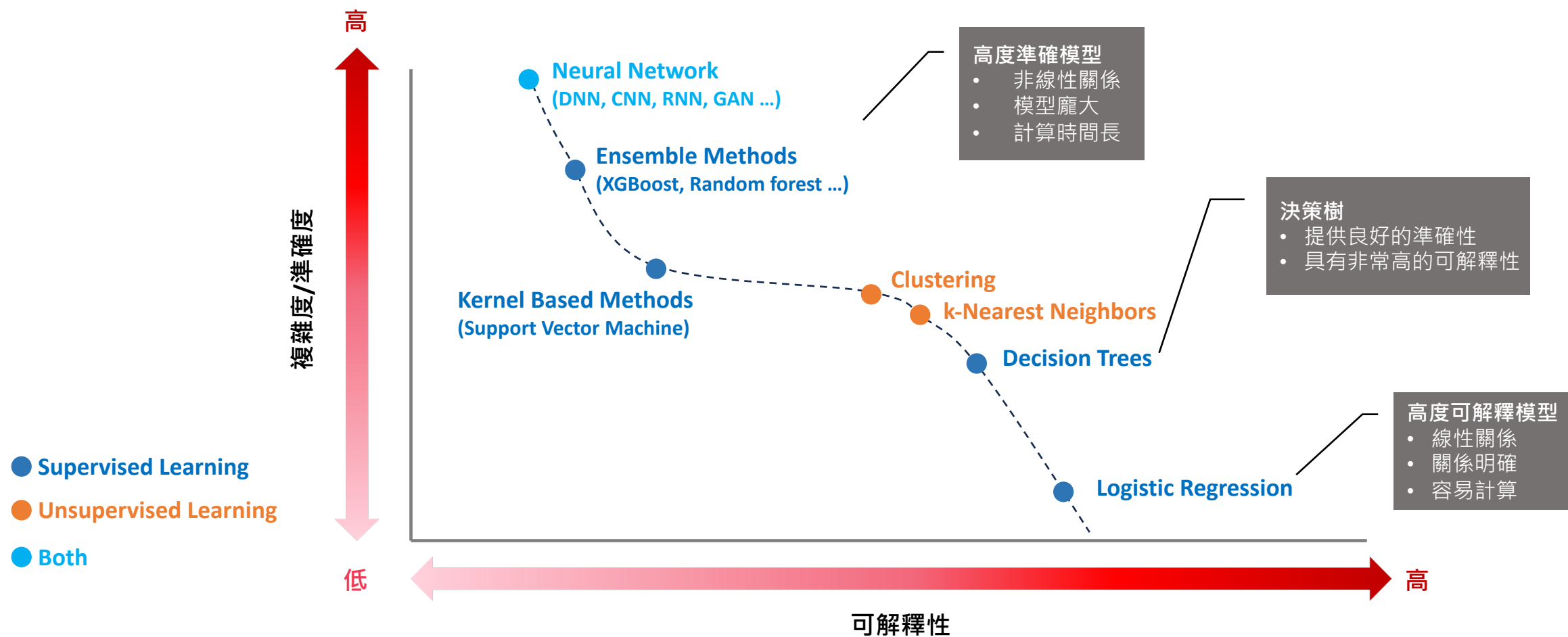
AI的推薦真的沒問題嗎？



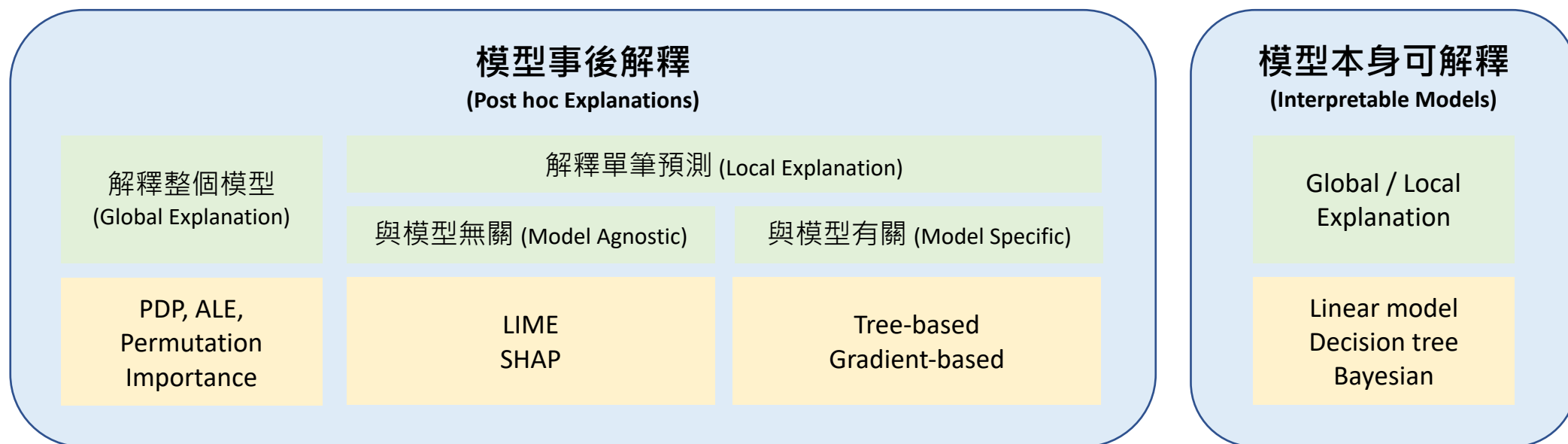
有了可解釋AI技術可以讓我們對AI模型更有信心！



準確度與可解釋性的權衡



模型可解釋性分成兩大類



第一類：模型本身可解釋

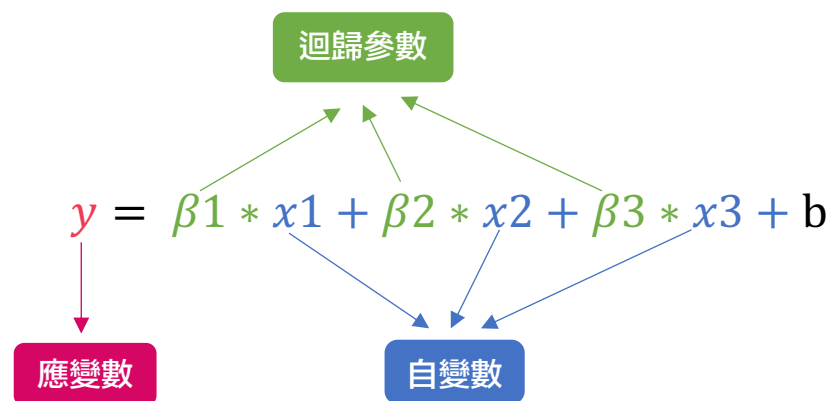
這一類型的ML模型可以視為具有解釋能力的「白箱」。

模型本身可解釋 (Interpretable Models)

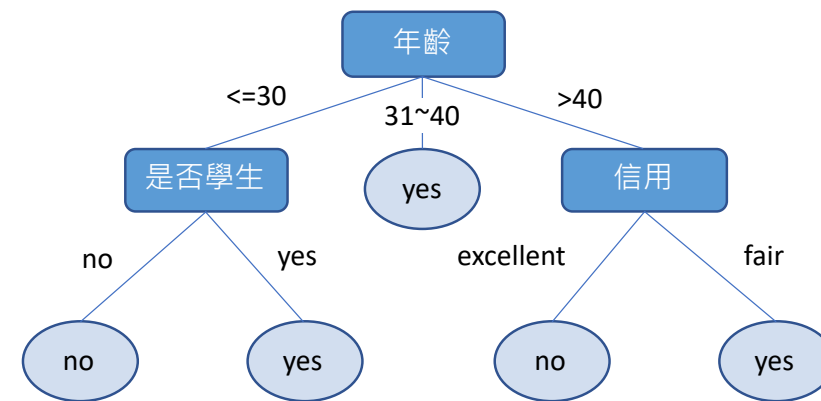
Global / Local
Explanation

Linear model
Decision tree
Bayesian

線性迴歸 (Linear Regression)



決策樹 (Decision tree)



第二類：模型事後解釋

當模型具有「黑箱」特性，因此必須透過外部方法解釋模型運作。

模型事後解釋 (Post hoc Explanations)

解釋整個模型
(Global Explanation)

PDP, ALE,
Permutation
Importance

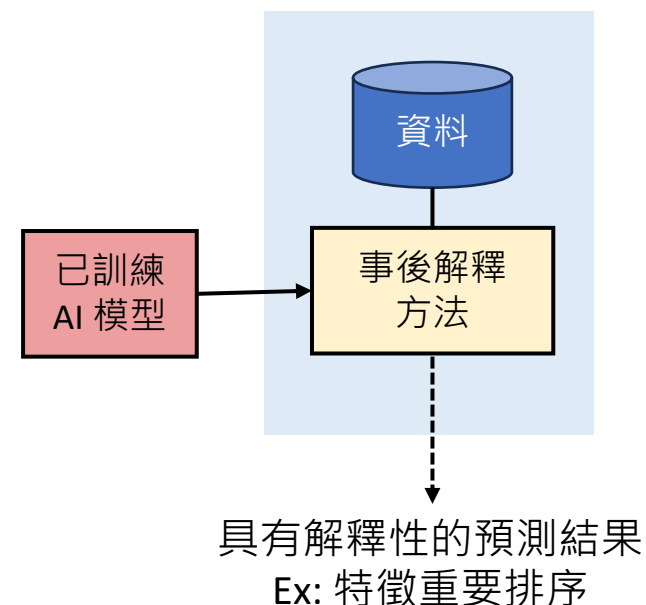
解釋單筆預測 (Local Explanation)

與模型無關 (Model Agnostic)

LIME
SHAP

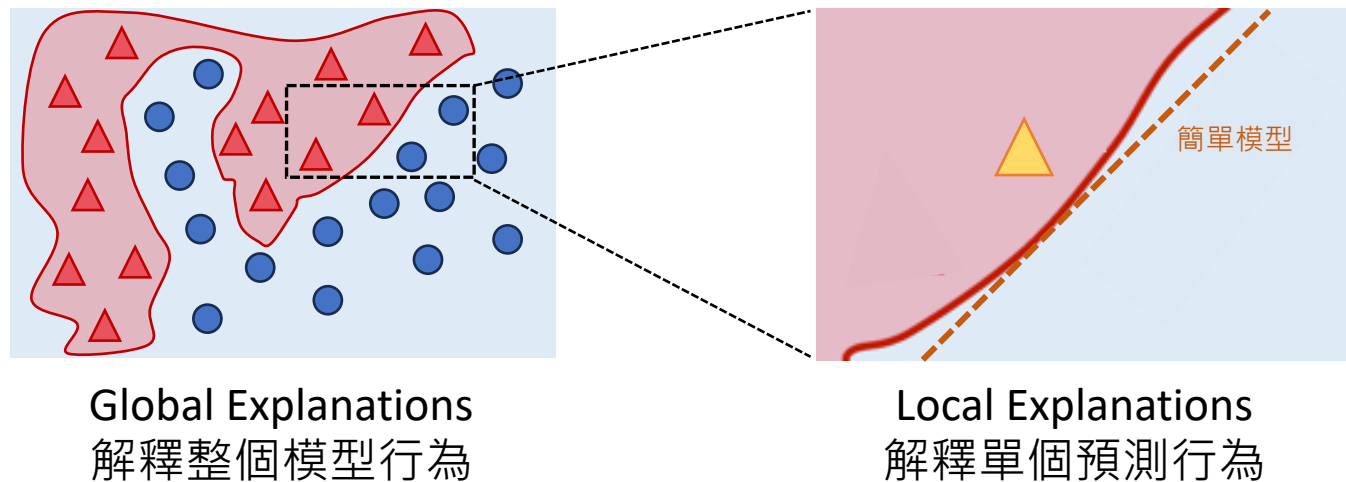
與模型有關 (Model Specific)

Tree-based
Gradient-based



SHAP vs. LIME

SHAP (SHapley Additive exPlanations)	LIME (Local Interpretable Model-agnostic Explanations)
與模型無關，透過資料解釋模型 (Model Agnostic)	
全局&局部解釋	局部解釋
用Shapely Value找貢獻值並解釋模型如何推論	透過建立簡單的模型解釋某筆資料

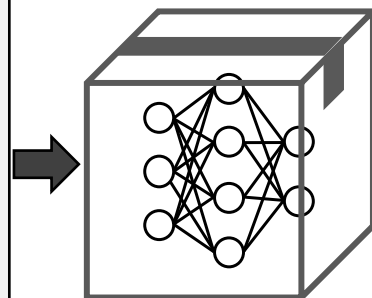


使用LIME解釋模型決策：「電信業」顧客流失預測

Input Data

Age 性別: Male
SeniorCitizen 是否長者: 0
Partner 配偶: No
Dependents 家人: No
Tenure 客戶停留月數: 1
PhoneService 服務: No
MultipleLines 多支門號: No
InternetService 網路: DSL
OnlineSecurity 網路安全服務: No
OnlineBackup 網路備份服務: No
DeviceProtection 保固: No
TechSupport 技術支援服務: No
StreamingTV 影音串流: No
StreamingMovies 電影: No
Contract 租約時間: Month-to-month
PaperlessBilling 電子帳單: No
PaymentMethod 付款方式: Bank transfer
MonthlyCharges 月租費: 24.8
TotalCharges 總花費: 24.8


Model (black box)

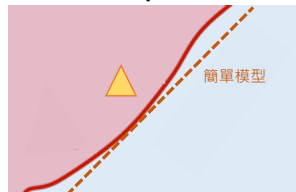


預測結果

No 0.37
Yes 0.63

客戶是否流失

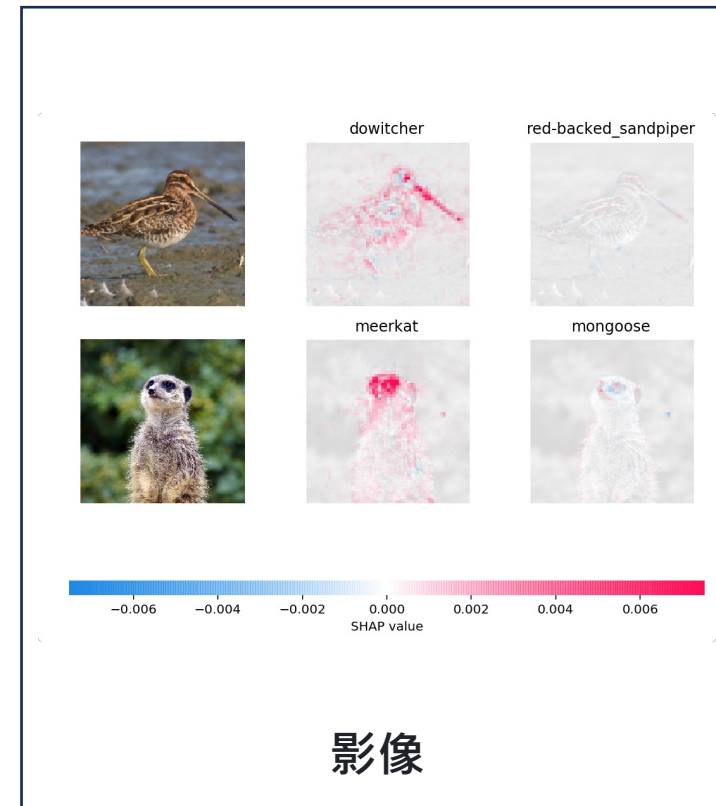
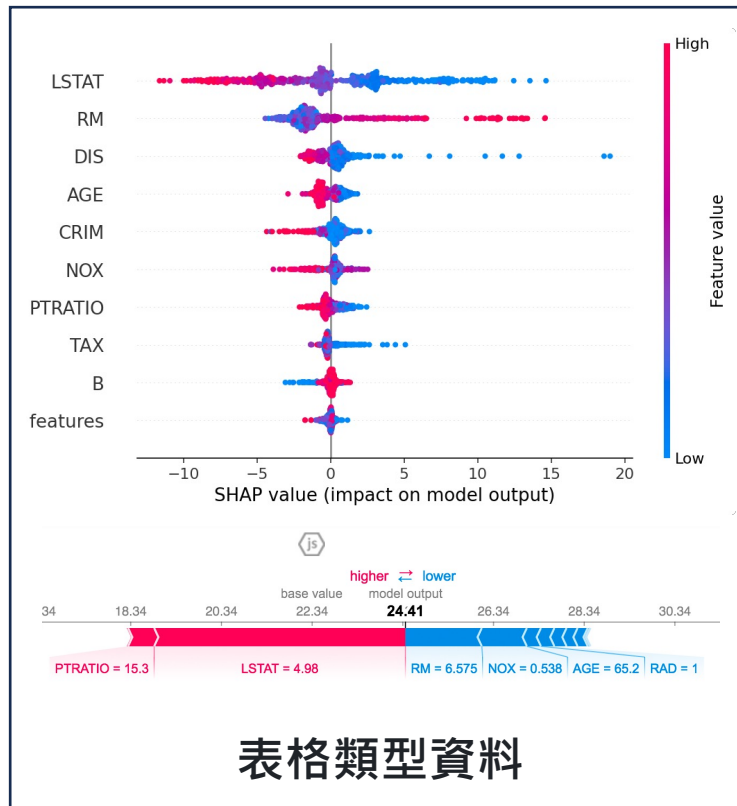
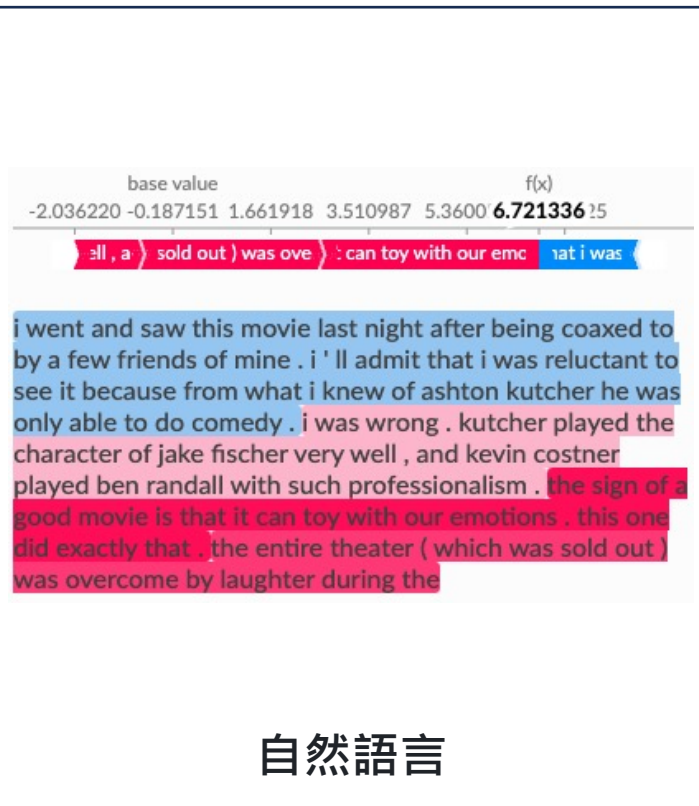

LIME
(Local Explanations)



Local explanation for class Yes



SHAP 在不同數據格式中的解釋性分析

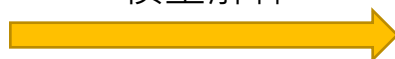


影像的可解釋技術

我們可以將神經網路的輸出結果視覺化，並觀察神經網路在推論過程中，哪些特徵影響最大。



模型解釋



重要特徵



Explainable CNN 技術

- Perturbation-based Explanation (擾動解釋)
- Gradient-based Explanation (梯度解釋)
- Propagation-based Explanation (傳播解釋)
- CAM-based Explanation (Class Activation Mapping 解釋)
- Attention-Based Explanation (基於注意力的解釋)



Perturbation-Based



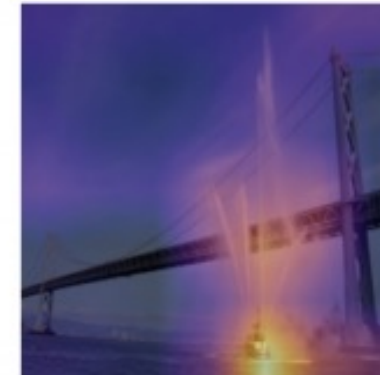
Gradient-Based



Propagation-Based



CAM-Based

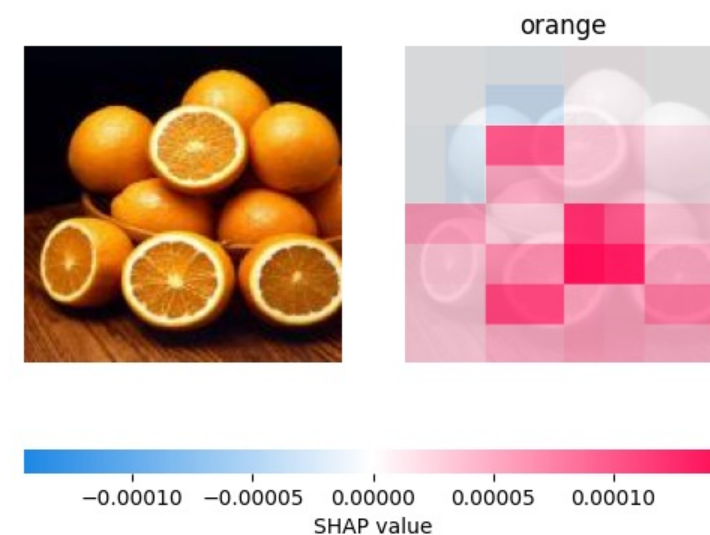
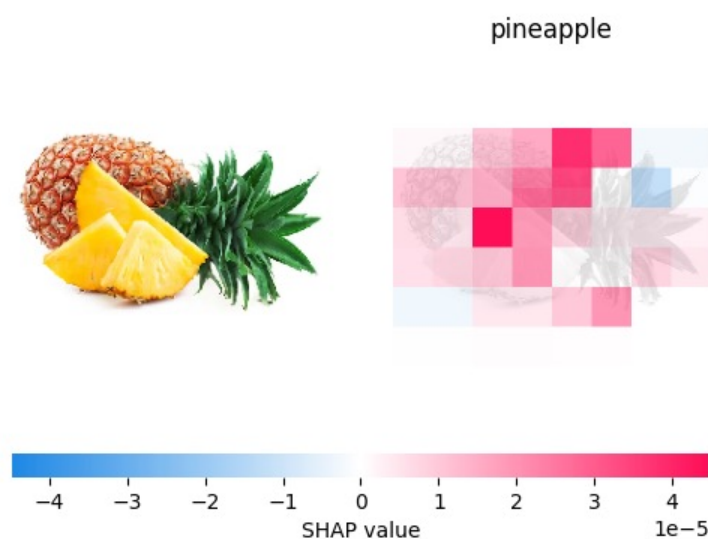
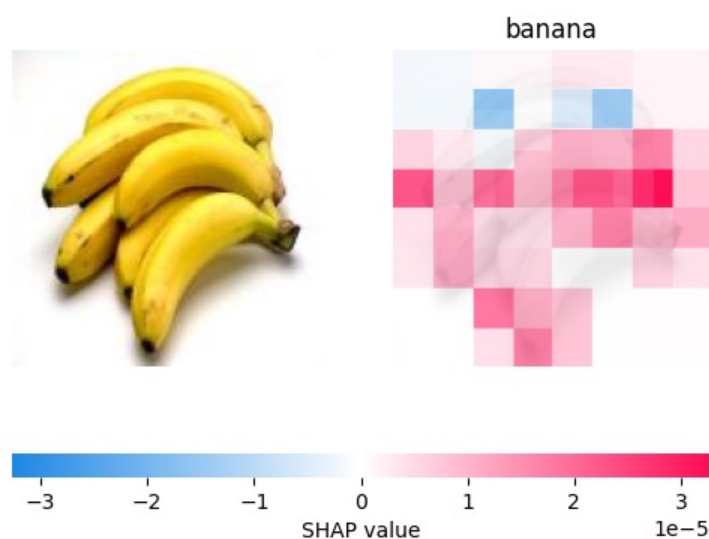


Attention-Based



使用 SHAP 解釋 CNN 模型

Partition Explainer 是 SHAP 套件中的一種方法，用於解釋機器學習模型。針對 CNN 模型的解釋，它可以用於分析圖像分類模型的決策。



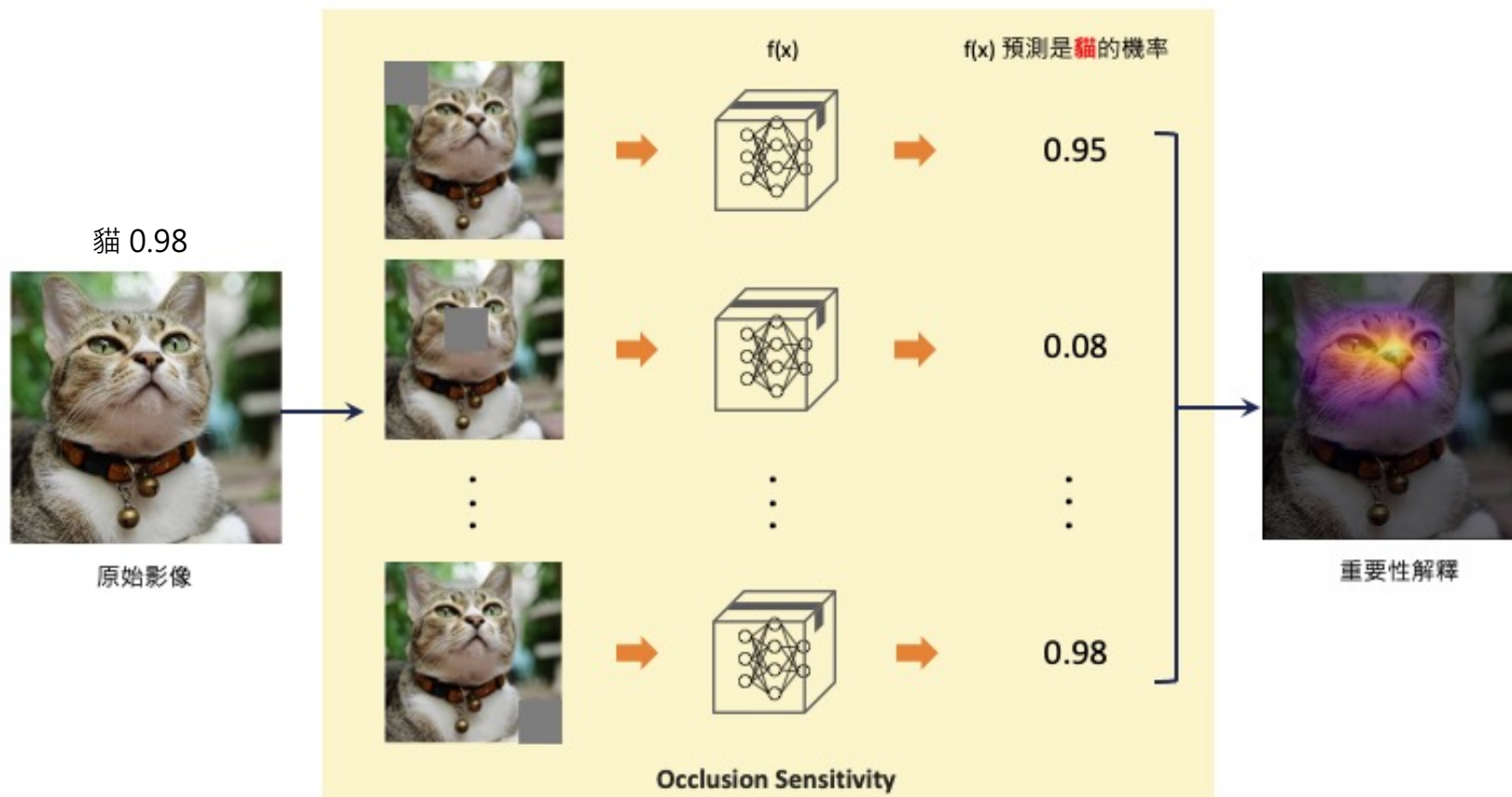
 Open in Colab

<https://reurl.cc/kyO4VG>



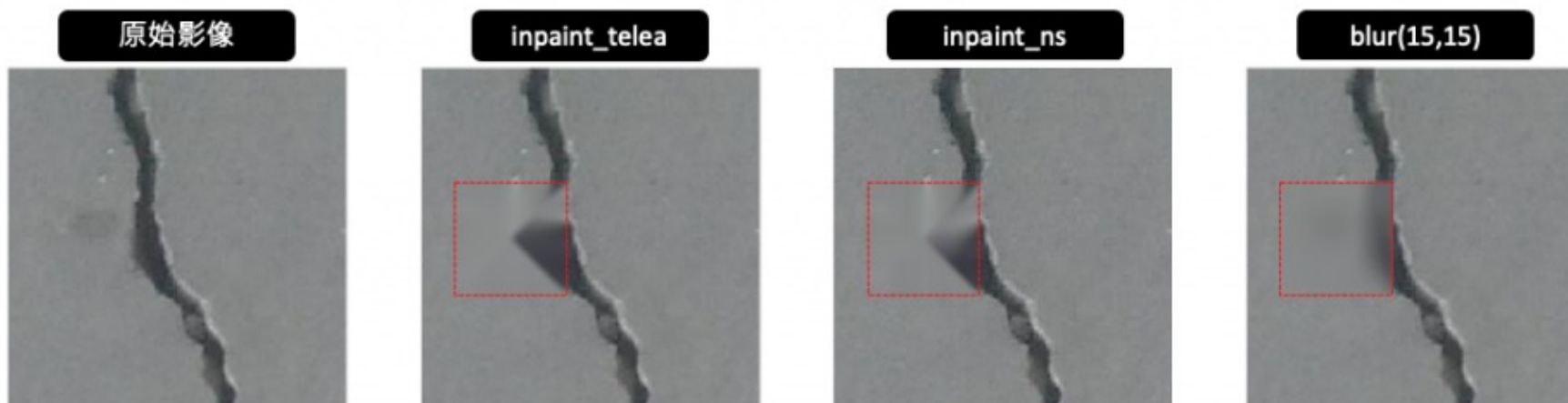
解釋運作原理

- Partition Explainer 是基於遮擋的方法



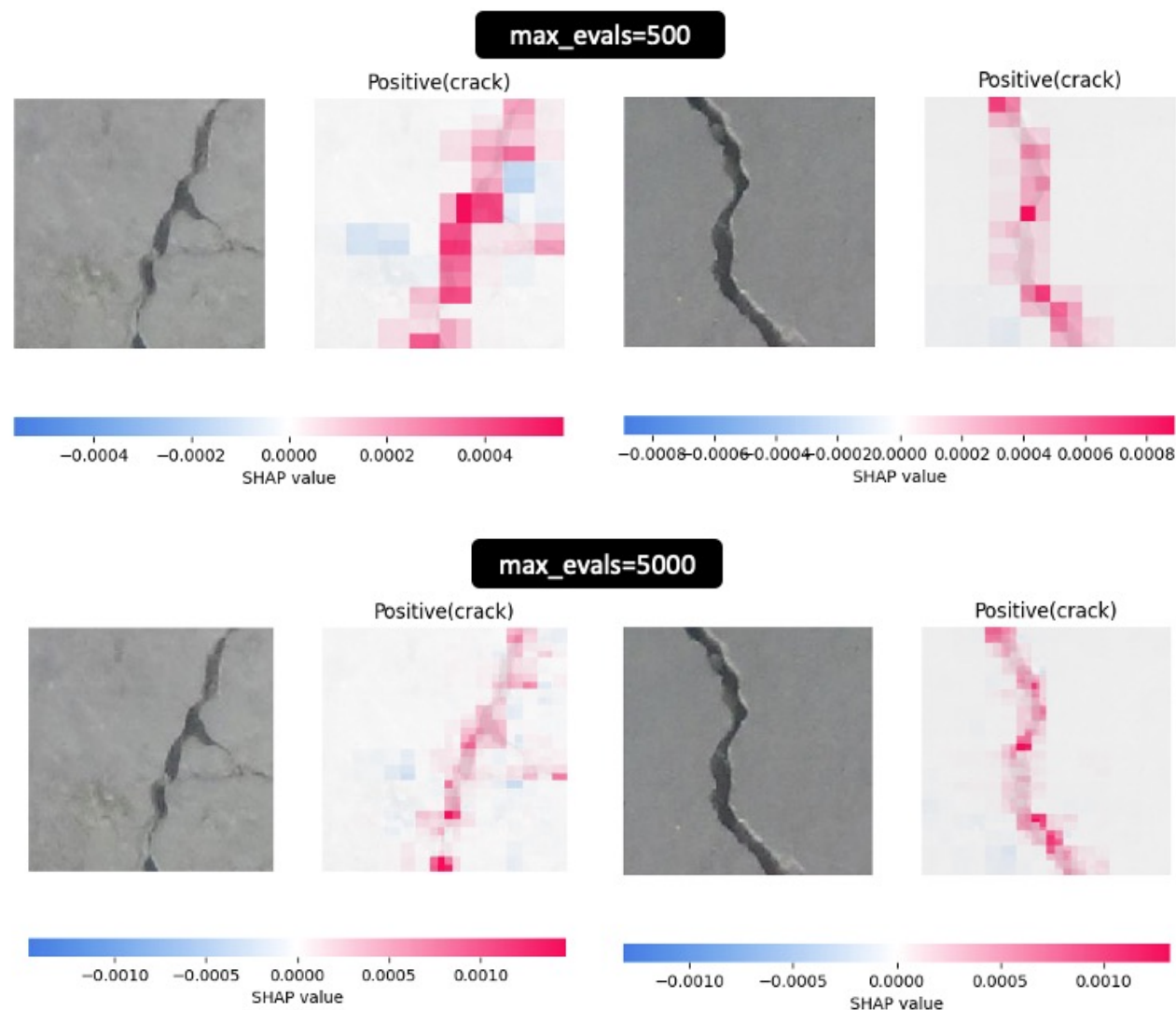
提供了三種方法來遮蓋圖像的部分區域

- `inpaint_telea`：使用 Telea 演算法，根據鄰近像素快速填補遮罩區域。
- `inpaint_ns`：使用 Navier-Stokes 演算法，適合填補具有複雜紋理的區域。
- `blur(kernel_xsize, kernel_ysize)`：使用指定大小的核來平均鄰域像素值，產生模糊效果。



Partition Explainer 中的 max_evals 參數

max_evals 參數代表在估算 SHAP 值時的最大評估次數。較大的值可能會提供更準確的 SHAP 值，但也會增加計算時間。



探索可解釋人工智慧 (免費學習資源)

2023 iT邦幫忙鐵人賽(AI & Data 組) [揭開黑箱模型：探索可解釋人工智慧](#)

全民瘋AI系列 [探索可解釋人工智慧]

搜尋

GitHub

全民瘋AI系列 [探索可解釋人工智慧]

1.XAI基礎與概念介紹 ^

[Day 1] 揭開模型的神秘面紗：為何XAI對機器學習如此重要？

[Day 2] 從黑盒到透明化：XAI技術的發展之路

[Day 3] 機器學習中的可解釋性指標

[Day 4] LIME vs. SHAP：哪種XAI解釋方法更適合你？

[Day 5] 淺談XAI與傳統機器學習的區別

2.XAI在傳統機器學習中的應用 ^

[Day 6] 非監督學習也能做到可解釋性？探索XAI在非監督學習中的應用

[Day 7] KNN與XAI：從鄰居中找出模型的決策邏輯

[Day 8] 解釋線性模型：探索線性迴歸和邏輯迴歸的可解釋性

[Day 9] 基於樹狀結構的XAI方法：決策樹的可解釋性

[Day 10] Permutation Importance：從特徵重要性角度解釋整個模型行為

[Day 11] Partial Dependence Plot：探索特徵對預測值的影響

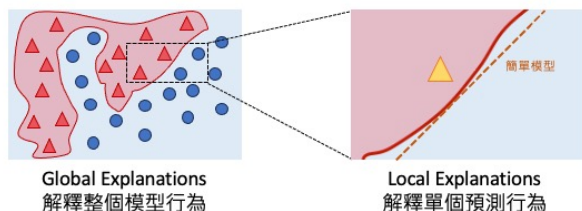
3.XAI常用工具介紹 ^

[Day 12] LIME理論：如何用局部解釋解釋模型行為

[Day 4] LIME vs. SHAP：哪種XAI解釋方法更適合你？

LIME 和 SHAP 都是機器學習中的解釋性方法，它們的共同點是都適用於模型無關性 (Model Agnostic)，並透過資料來解釋模型的預測結果。如果不想深入模型內部，但又想要能解釋模型，那看這篇文章就對了！簡單來說今天介紹的解釋方法不探究模型內部的運作原理，而是嘗試帶一些資料，去觀察輸出結果。然而 LIME 與 SHAP 在解釋方式和範圍上有一些區別。在今天的文章中，我們不會深入探究這兩個方法的背後詳細原理。而是透過一個簡單地例子，讓各位讀者知道兩種方法的差別與使用時機。

LIME	SHAP
透過資料解釋模型 (Model Agnostic)	
局部解釋	全局&局部解釋
透過建立簡單的模型解釋某筆資料。	用 Shapely Value 找貢獻值並解釋模型如何推論。



本頁目錄

LIME

LIME 的局部解釋過程

LIME 應用例 (表格資料)

SHAP

SHAP 的全局/局部解釋過程

SHAP 應用例 (表格資料)

小結

Reference

• 2023 iThome 鐵人賽

[回列表](#)

AI & Data

揭開黑箱模型：探索可解釋人工智慧 系列

本系列將從 XAI 的基礎知識出發，深入探討可解釋人工智慧在機器學習和深度學習中的應用、案例和挑戰，以及未來發展方向。希望透過這個系列，幫助讀者更好地理解應用可解釋人工智慧技術，促進可信、透明、負責任

[展開](#)

鐵人鍊成 | 共 30 篇文章 | 22 人訂閱

[RSS系列文](#)

0

Like

0

留言

1046

瀏覽

DAY 1

[Day 1] 揭開模型的神秘面紗：為何XAI對機器學習如此重要？

人工智慧的發展已經進入了一個新的階段，作為AI的重要分支「機器學習」，已經被廣泛應用於各個領域，例如語音識別、圖像分類、自然語言處理.....等。然而隨著機器...

2023-09-14 · 由 [10程式中](#) 分享

1

Like

0

留言

1436

瀏覽

DAY 2

[Day 2] 從黑盒到透明化：XAI技術的發展之路

近年來人工智慧技術發展迅速，深度學習等技術的出現和應用已經帶來了很多驚人的成果，尤其是 ChatGPT 的出現更讓人們驚嘆不已。然而這些模型的黑箱特性一直是人工...

2023-09-15 · 由 [10程式中](#) 分享

0

Like

0

留言

1446

瀏覽

DAY 3

[Day 3] 機器學習中的可解釋性指標

「可解釋性指標」是 XAI 中用來衡量模型可解釋性的評估標準。它們是用來確定模型如何解釋其預測的方式，以及如何在給定輸入後生成可解釋的結果。可解釋性指標可以根據...

2023-09-16 · 由 [10程式中](#) 分享

全民瘋AI系列免費電子書上線囉！

<https://andy6804tw.github.io/crazyai-xai>

